When Measurement Errors Correlate with Truth: Surprising Effects of Nondifferential Misclassification

Sholom Wacholder

Most of the literature on the effect of nondifferential misclassification and errors in variables either addresses binary exposure variables or discusses continuous variables in the classical error model, where the error is assumed to be uncorrelated with the true value. In both of these situations, an imperfectly measured exposure always attenuates the relation, at least in the univariate setting. Furthermore, measuring a confounder with error independent of the exposure, even while measuring the exposure of interest perfectly, leads to partial control of the confounding. For many variables measured in epidemiology, particularly those based on self-report, however, errors are often correlated with the true value, and these rules may not apply. Epidemiologists need to be wary of deviations from the classical error model, since poor measurement might occasionally explain a positive finding even when the error does not differ by disease status. (Epidemiology 1995;6:157–161)

Keywords: bias, biometry, environmental exposure, epidemiologic methods, occupational exposure, odds ratio, sensitivity and specificity, regression analysis, statistics.

Epidemiologists often must use data measured with error. Sometimes we can rely on statistical demonstrations that the errors will affect parameter estimates in predictable ways.[1,2] In particular, estimates of relative risk will be biased toward but not beyond the null for binary exposures and for continuous variables under the classical error model, in which errors are assumed to be independent of the true value. In this paper, I discuss a more general error model that encompasses the situation of errors related to the true value. I present the conditions under which estimates of effect or regression estimates can be exaggerated or can reverse direction. By considering the more general situation, one can reconcile some recent work with some of the earlier epidemiology literature in this area.

Error Models

Classical Error Model

The classical error model assumes that the magnitude and direction of the error in measuring a variable do not depend on its true value. Thus, the errors in measuring large and small values will have the same average value, for example. In the classical error model, the observed variable Z is related to the true value X according to:

$$Z = X + E \qquad (1)$$

where E, the error in measuring X, is assumed to have mean zero and be independent of X. That is, the direction and magnitude of the errors of measurement are not related to the actual value of X. This model is realistic in many circumstances, such as when errors are the consequence of sloppy laboratory technique or difficulty in using an instrument. Under the classical error model, the slope of the regression of Z on X would be 1 and the intercept 0.

An Alternative to the Classical Error Model

In many common situations in epidemiology, the classical error model does not hold. When Z is a self-reported value, it seems unrealistic to expect errors to be independent of true values. For example, errors in self-reported height and weight seldom follow model 1, since the errors do not have mean zero and are correlated with the true value.[3,4] Also, it is possible that those who eat smaller amounts of a nutrient may tend to overreport, whereas those who eat larger amounts may underreport consumption.[5]

A more general error model[6] that allows E to be correlated with X needs to be considered. Let $\sigma^2 \equiv \text{var}(X)$, $\omega^2 \equiv \text{var}(E)$, $\phi \equiv \text{cov}(X,E)$, and the correlation of X and E be $\rho = \phi/(\sigma\omega)$. The bias factor in this error model, given by equation 8.8 of Cochran,[6] is the ratio of the slope γ of the regression of dependent variable Y on the observed value of the independent variable Z relative to the slope β of the regression of Y on the true values of the independent variable X:

$$BIAS = \frac{\gamma}{\beta} = \frac{\sigma^2 + \phi}{\sigma^2 + 2\phi + \omega^2} \qquad (2)$$

$$= \frac{\sigma^2 + \rho\sigma\omega}{\sigma^2 + 2\rho\sigma\omega + \omega^2}$$

From the Biostatistics Branch, National Cancer Institute, 6130 Executive Boulevard, EPN 403, Rockville, MD 20852.

Submitted March 14, 1994; final version accepted June 17, 1994.

The special case of $\rho = 0$ is the classical error model, in which expression 2 reduces to the attenuation factor[2]:

$$BIAS = \frac{\sigma^2}{\sigma^2 + \omega^2} \tag{3}$$

bounded between 0 and 1. In the more general model 2, the bias factor can be negative or greater than one. So perhaps the term "attenuation factor" is misleading, and "distortion factor" better reflects the more general situation.

Implications of the General Error Model for Univariate Regression

Exaggeration and reversal of regression coefficients are both possible. Table 1 displays the distortion factor from Eq 2 as a function of the ratio σ^2/ω^2 and ρ, the correlation between X and E. Simple algebraic manipulation of expression 2 confirms the suggestion of the table that $\gamma > \beta$, that is, there is exaggeration or deattenuation of the regression effect, when $\phi < -\omega^2$, or equivalently, when $\rho < -\omega/\sigma$. Thus, when $\omega^2 < \sigma^2$, that is, the error variance is less than the variance of X, a strong negative correlation between X and E can result in bias effects that are not possible under the classical error model. When $\omega^2 > \sigma^2$, exaggeration cannot occur since ρ is always greater than -1. Even when $\rho < -\omega/\sigma$, there cannot be substantial exaggeration, as shown in Table 1.

Reversal of direction of effect can occur when $\phi < -\sigma^2$, or equivalently, $\rho < -\sigma/\omega$. Thus, reversal requires that the error variance be greater than the population variance, as well as a strong negative correlation between X and E.

When $\rho = -\omega/\sigma$, there is no bias from estimating γ instead of β. This phenomenon is equivalent to the Berkson error model,[2,7–9] in which E is uncorrelated with Z, rather than E uncorrelated with X, as in the classical error model. The distortion factor from Eq 2 is 1 in the Berkson model [since $0 = \text{cov}(Z,E) = \text{cov}(X + E,E) = \phi + \omega^2$], and, therefore, there is no bias in estimating β. Berkson[7] gave the example of a bioassay of a material assigned to have dose level Z (in our notation) but actually receiving level X, where it seems reasonable to assume that the measurement error $E = Z - X$ is independent of Z.

This approach is useful even when there is bias in Z, that is, the mean of E is not zero. We illustrate the point with the error model $Z = a + sX$ for intercept a and slope s; for simplicity, no stochastic element is allowed in Z. Clearly, a regression model with coefficient $\beta > 0$ for X will have coefficient $\gamma = \beta/s$ for Z. Here, $E = Z - X = a + (1 - s)X$, so, assuming, without loss of generality, $\sigma^2 = 1$ implies $\omega^2 = (1 - s)^2$, $\rho = -1$ and $\phi = -(1 - s)$. Thus, $s > 1$ implies $\gamma < \beta$; $s < 1$ implies $\gamma > \beta$; and $s < 0$ implies $\gamma < 0 < \beta$. A special case, resulting from catastrophic error, is $Z = -X$, ($a = 0$ and $s = -1$), where the bias factor is -1.

Plots

The possible impact of different forms of error on linear regression estimates is demonstrated graphically in Figure 1, using hypothetical data found in Table 2. In this simple example, $\sigma^2 = \text{var}(X)$ is 2, the slope β is 1, and $\text{var}(Y \mid X)$ is 0. The *solid line* in each plot is the regression of Y on the true values of X. The *broken line* in each plot displays the regression of Y on $Z = X + E$ together with the points (Y,Z) for a specified error structure. In the classical error model, E is uncorrelated with X, and, so, by Eq 3, the slope of Y on Z is attenuated by the factor $1/(1 + \omega^2/\sigma^2) = 1/[1 + \text{var}(E_1)/\text{var}(X)] = 0.87$. In the plot with the Berkson error model, $\phi = -0.8$, $\omega^2 = 0.8$, and $\text{cov}(E_2,Z_2) = \text{cov}(E_2,X + E_2) = \text{cov}(E_2,X) + \text{var}(E_2) = 0$, leading to a bias factor of $(2 - 1.6)/(2 - 2 \cdot 1.6 + 1.6) = 1$; that is, no bias. When the error is negatively $[\text{cov}(X,E) = \phi = -1.6]$ or positively ($\phi = +1.6$) correlated with X, the bias factors for the slope can be calculated from Eq 2, using $\omega^2 = 1.3$, as $(2 - 1.6)/(2 - 2 \cdot 1.6 + 1.3) = 4$ and $(2 + 1.6)/(2 + 2 \cdot 1.6 + 1.3) = 0.55$.

These results follow the theory described in earlier subsections. For error independent of or positively correlated with the true value, the slopes are attenuated, as always. For error negatively correlated with the true value, the slope can be exaggerated or unchanged as shown in the figure, attenuated when there is a weak negative correlation between X and E, or reversed in direction when $\omega^2 = \text{var}(E)$ is much greater than $\text{var}(X) = \sigma^2$.

TABLE 1. Bias Factor as a Function of var(E)/var(X) = ω^2/σ^2 and corr(X, E) = ρ (Based on Eq 2)

	ρ								
ω^2/σ^2	−1	−0.75	−0.5	−0.25	0	0.25	0.5	0.75	1
0.01	1.1	1.1	1.04	1.02	0.99	0.97	0.95	0.93	0.91
0.1	1.5	1.2	1.07	0.98	0.91	0.86	0.82	0.79	0.76
0.5	3.4	1.1	0.82	0.72	0.67	0.64	0.61	0.60	0.59
0.8	9.5	0.72	0.61	0.57	0.56	0.54	0.54	0.53	0.53
1		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
1.25	−8.5	0.28	0.39	0.43	0.44	0.46	0.46	0.47	0.47
2	−2.4	−0.069	0.18	0.28	0.33	0.365	0.39	0.40	0.41
10	−0.46	−0.22	−0.074	0.022	0.091	0.14	0.18	0.21	0.24
50	−0.16	−0.11	−0.058	−0.016	0.020	0.051	0.078	0.10	0.12
100	−0.11	−0.076	−0.044	−0.016	0.0099	0.033	0.054	0.073	0.091

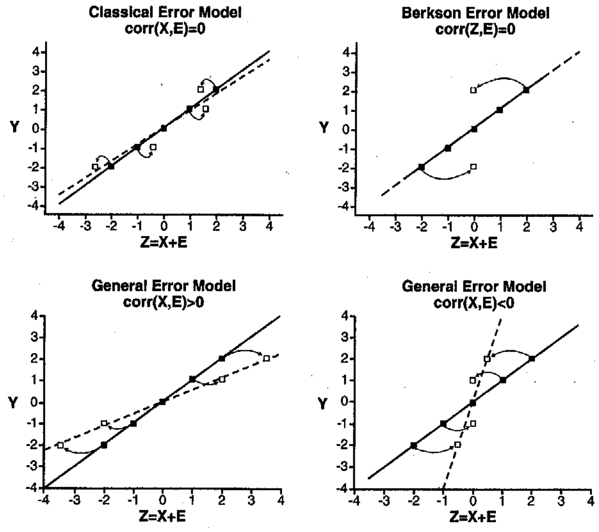

FIGURE 1. **Plot of points and regression lines from hypothetical true and observed data for several error structures. In each plot, the points (X,Y) and (Z,Y) are represented by a *closed* and an *open square*, respectively. The *solid* and *broken lines* are the regressions of Y on X and on Z, respectively. In the Berkson error model, the two regression lines are identical.**

TABLE 2. Hypothetical Data Used in Plots

		E			
	$X = Y$	Classical	Berkson	Negative Correlation	Positive Correlation
	−2	−0.6	2.0	−1.5	1.5
	−1	0.6	0.0	−1.0	1.0
	0	0	0.0	0	0
	1	0.6	0.0	1.0	−1.0
	2	−0.6	−2.0	1.5	−1.5
$\phi = \text{cov}(X, E)$		0	−1.6	1.6	−1.6
Variance	2	0.29	1.6	1.3	1.3

Binary Regression Variables

Misclassified binary variables always have been treated as a separate case, since the classical error model 1 clearly does not apply. Nonetheless, the impact of the misclassification is also attenuation of the effect (except in extreme conditions). Algebra in the appendix shows that the well-known results for binary X are simply a special case of the general error model 2.

Trends in Polytomous Regression Variables

The estimate of trend for a polytomous exposure variable is really an estimate of the slope of the regression on a continuous variable. Thus, one can easily demonstrate the possibility of an exaggerated or reversed trend from analysis of epidemiologic data with a misclassified polytomous exposure. The left panel of Table 2 in the study by Dosemeci *et al*[10] is an example with logistic analysis of case-control data. A similar example for a cohort study analyzed by Poisson regression can be constructed easily. Of course, when there are only two levels of exposure, the binary results described in the previous subsection apply.

Summary of Univariate Results

One can see that it is possible to have exaggeration or reversal of the slope of a univariate regression on a continuous variable or in a univariate logistic regression even when the errors are nondifferential. The general formulation encompasses binary variables and other important special cases not considered by the classical error model.

Negative correlation between the error and the true value for exposure will occur in many situations: for example, when the misclassification at higher levels of exposure tends to be toward a lower level, whereas there is either little misclassification at lower levels of exposure or the misclassification at lower levels tends to be toward a higher level.

In a multivariate setting, these results apply directly to the slope of the regression of Y on X (with variances and covariances made conditional on other variables in the model), as long as there is no error in the other covariables. A more general formulation of this errors-in-variables model would account for the possibility of correlated errors among several covariates, thereby including polytomous exposure variables. In the next section, the effect of errors in a confounding variable on the adjusted estimate of effect of an exposure variable measured without error is considered.

Implications for Confounding Variables

Several articles in the literature[2,9,11–14] have claimed that when a confounding variable is measured with error that is independent of both exposure and disease, whereas the exposure variable is measured without error, partial control of the confounder is achieved, that is, the adjusted estimate lies between the crude estimate and the estimate that would have been obtained had the confounder been measured without error. Although true when the confounder is binary, Brenner[15] recently demonstrated that the claim is false for a polytomous confounder. I will show that the claim holds for linear regression under the classical error model but not when errors are correlated to the true value of a continuous confounder. Consider the regression model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \qquad (4)$$

One property of confounding is that

$$\beta_1 = \beta_1^* + S_{21}\beta_2, \qquad (5)$$

where β_1^* is the slope from the crude regression of Y on X_1, and S_{21} is the slope of the regression of X_2 on X_1 (equation 17.11.1 in Snedecor and Cochran's text[8]). If Z_2 is the observed value of X_2, the regression model

$$E(Y) = \gamma_0 + \gamma_1 X_1 + \gamma_2 Z_2 \qquad (6)$$

is fit instead of model 4. If we observe $Z_2 = X_2 + E_2$ but $Z_1 = X_1$ is observed without error, it is possible that either $\gamma_2 > \beta_2 > 0$ [when E_2 is nondifferential for Y and $\text{cov}(E_2,X_2) < -\text{var}(X_2)$], or that $\gamma_2 < 0 < \beta_2$, as noted in the previous section. If $\gamma_2 > \beta_2 > 0$; the slopes of the regressions of X_2 and Z_2 on X_1 are equal; and E_2 is nondifferential for Y, then $\gamma_1 = \beta_1^* + S_{21}\gamma_2 > \beta_1^* + S_{21}\beta_2 = \beta_1 > \beta_1^*$. That is, the estimate adjusted for the covariate measured with error does not lie between the crude and the correctly adjusted values. If the classical error model holds, $\beta_2 > \gamma_2 > 0$, and the adjusted estimate would indeed be bounded.

When $\text{cov}(E_2,Z_2) = 0$, that is, in the Berkson error case, it can be shown that corresponding regression coefficients in Eqs 4 and 6 are equal if the errors E_2 in X_2 are independent of X_1, conditional on X_2. The two factors in the correction term in Eq 5 are unchanged: β_2 by the Berkson error assumption, and S_{21} by the assumption that E_2 is independent of X_1, conditional on X_2.

Thus, the theory for nondifferential error in the confounder independent of exposure parallels that of the univariate case. Errors fitting the Berkson model produce no bias. Errors in a binary confounder or errors that follow the classical error model for a continuous confounder predictably control partially for confounding, so it is better to adjust for a misclassified confounder than not to adjust. But errors in a polytomous confounder or errors correlated with the true value of a continuous confounder may produce unpredictable bias. Indeed, errors that strongly correlate with the true value of the confounder or with the exposure can produce the apparent anomaly that adjustment for a poorly measured variable yields an estimate that is more biased than the crude.

Discussion

Interpretation of epidemiologic findings can rely securely upon the well-described attenuation of effect resulting from nondifferential error in binary variables or in continuous variables when the error is uncorrelated to the true value. But epidemiologists need to be aware that attenuation will not be assured when: (1) error is differential; (2) independent variables are polytomous; or (3) independent variables are continuous with errors correlated with the true value.

Thus, epidemiologists should not be too quick to assert automatically that poor measurement cannot explain a positive finding. Furthermore, study designs that tolerate unnecessary errors in one group so that errors are not differential should be examined carefully. Strict adherence to the principle of comparable accuracy used to ensure nondifferential misclassification in choosing controls for case-control studies may not be advisable when it would require controls with as much error as cases instead of more accurate controls.[16]

Freedman *et al*[17] discuss the effects of dietary measurement error on sample size requirements under the classical error model assumption. Extensions of that work might explicitly quantify the effect of the relation between errors and the true values on study size; for example, negative correlation between the true value and the reported value would result in a smaller study size than when the values are uncorrelated.[17]

The classical error model assumption (error is independent of the true value) is often reasonable when errors are entirely due to the observer, but may be unrealistic when information is gathered from self-reports. A tendency for respondents to give answers close to socially acceptable norms would lead to violation of the assumption; a person with a much higher than average value would tend to give an answer below the true value, whereas someone with a low value might tend to reply with a positive error, resulting in the "flattened-slope syndrome."[5] Thus, there would be a negative correlation between the error and the true value unless the errors for those with below average fat consumption tend to be underreported even more.

I do not wish to overemphasize the frequency or importance of exaggeration or reversal of the estimate of effect of a continuous variable or of trend (Table 1). Substantial exaggeration is probably rare, since a strong negative correlation (say, $\rho < -0.5$) between error and truth is required. Reversal of direction requires a strong negative correlation and that the variance of the error be greater than the variance of the true value. On the other hand, the upper right-hand portion of Table 1, with positive correlation and less variance in E than in X, seems realistic in some situations where measurements are related; there we see moderately more attenuation than would be predicted by Eq 3, that is, by assuming $\rho = 0$. Since it is difficult to assess the error structure or observe ρ, ω^2, or σ^2 except in the rare situation when a gold standard is available,[18] one cannot usually rely on data to quantify the extent of bias precisely and must often use information from outside the study to evaluate the nature of the bias.

Considerations of error structure are important not only in interpreting individual studies but also for methods of correcting for error. Some correction procedures based on a validation study may not be robust against departures from assumed error models. The incorrect assumption of a classical error model, just like that of nondifferential misclassification or of the infallibility of an "alloyed" gold standard,[18] can make the results of a validation study misleading.

Rothman[19] has suggested using nondifferential misclassification theory in choosing among several hypothesized formulations of exposure. He argues that "the closer the assumption is to the truth, the larger will be the measured effect."[19,p58] But under the general error model, regression on the incorrect form Z can result in a higher slope and a stronger apparent association than regression on X, so this procedure is not robust against errors that do not follow the classical model. Indeed, even when a continuous exposure is measured with nondifferential error, categorization can induce differential misclassification and, hence, exaggeration of effect.[20,21]

Consequences of errors in confounders can be serious, even when nondifferential.[11,22,23] Adjustment for poorly measured polytomous[15] or continuous confounders, even when error or misclassification is independent of exposure and disease, can result in a poorer estimate of the adjusted effect than would be obtained by relying directly on the crude estimate. Some consideration should be given to identifying when adjustment for poorly measured variables like socioeconomic status is likely to be more biased than not adjusting at all. A similar point holds for adjustment for total energy in studies of macronutrients, particularly since the errors in the exposure and confounder are likely to be strongly correlated.

Epidemiologists have been misled by the emphasis on the classical error model in the literature and by the superficially similar result regarding the impact of misclassification of binary variables in a univariate model. I have shown the danger of reliance on the binary variable case or the classical error model. The more general error model presented here illuminates some correct but seemingly anomalous results showing reversal of direction in trend or at some levels of a polytomous variable despite nondifferential misclassification[10,24–28] and may help to limit claims[9,11–14] that are not true in the general case. Together with empirical investigations of error models and mechanisms, admittedly daunting areas for research given the scarcity of gold standards, it can also lead to a more realistic understanding of the impact of imperfect measurement of important epidemiologic variables.

References

1. Bross I. Misclassification in 2 × 2 tables. Biometrics 1954;10:478–486.
2. Armstrong BG. The effects of measurement error on relative risk regression. Am J Epidemiol 1990;132:1176–1184.
3. Palta M, Prineas RJ, Berman R, Hannan P. Comparison of self-reported and measured height and weight. Am J Epidemiol 1982;115:223–230.
4. Stewart AW, Jackson RT, Ford MA, Beaglehole R. Underestimation of relative weight by use of self-reported height and weight. Am J Epidemiol 1987;125:122–126.
5. Anonymous. Nutrient adequacy. Washington DC: National Academy Press, 1986;48–65.
6. Cochran WG. Errors of measurement in statistics. Technometrics 1968;10: 637–666.
7. Berkson J. Are there two regressions? J Am Stat Assoc 1950;45:164–180.
8. Snedecor GW, Cochran WG. Statistical Methods. 7th ed. Ames, IA: Iowa State University Press, 1980.
9. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 1993;14:69–93.
10. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification always bias a true effect toward the null value? Am J Epidemiol 1990;132: 746–748.
11. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol 1980;112:564–569.
12. Savitz DA, Baron AE. Estimating and correcting for confounder misclassification. Am J Epidemiol 1989;129:1062–1071.
13. Alderman BW, Baron AE, Savitz DA. Cautions in the use of antecedents as surrogates for confounders. Am J Epidemiol 1993;137:1259–1272.
14. Mertens TE. Estimating the effects of misclassification. Lancet 1993;342: 418–421.
15. Brenner H. Bias due to non-differential misclassification of polytomous confounders. J Clin Epidemiol 1993;46:57–63.
16. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. Am J Epidemiol 1992;135: 1019–1028.
17. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990;132:1185–1195.
18. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. Am J Epidemiol 1992;137:1251–1258.
19. Rothman KJ. Modern Epidemiology. Boston: Little, Brown, 1986;58.
20. Wacholder S, Dosemeci M, Lubin J. Blind assignment of exposure does not always prevent nondifferential misclassification. Am J Epidemiol 1991;134: 433–437.
21. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991;134: 1233–1244.
22. Walker AM. Misclassified confounders (Letter). Am J Epidemiol 1985;122: 921–922.
23. Ahlbom A, Steinbeck G. Aspects of misclassification of confounding factors. Am J Ind Med 1992;21:107–112.
24. Dosemeci M, Wacholder S, Lubin JH. The authors clarify and respond (Letter). Am J Epidemiol 1991;134:441–442.
25. Gilbert ES. Re: "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" (Letter). Am J Epidemiol 1991; 134:440–441.
26. Brenner H. Re: "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" (Letter). Am J Epidemiol 1991; 134:438–439.
27. Dosemeci M, Wacholder S, Lubin JH. Re: "Does nondifferential misclassification of exposure always bias a true effect toward the null value? The authors reply" (Letter). Am J Epidemiol 1992;135:1430–1431.
28. Myers J, Erlich R. Re: "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" (Letter). Am J Epidemiol 1992;135:1429–1430.
29. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978;107:71–76.

Appendix

In this appendix, the general error model results are applied for a binary variable X. If $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$ and sensitivity s and specificity t, $\mathrm{var}(X) = \sigma^2 = p(1 - p)$, $\mathrm{var}(E) = \omega^2 = -p^2(t + s - 2)^2 + p(2t^2 - 5t - 3s + 2st + 4) + t - t^2$ and $\mathrm{cov}(X,E) = \phi = -p(1 - p)(2 - s - t)$. Comparison of σ^2 and ϕ reveals that there is no apparent effect when the sum of sensitivity and specificity equals one ($s + t = 1$), and reversal of direction when $s + t < 1$, as noted often before.[1,6,29] Since an increase in E results in an increase in Z, $0 < \mathrm{cov}(Z,E) = \mathrm{cov}(X,E) + \mathrm{var}(E)$, or $\phi > -\omega^2$ and $\rho > -\sigma/\omega$. Thus, by the results above, there can be no exaggeration for misclassification of binary variables.